Using Exploration and Learning for Medical Records Search: An Experiment in Identifying Cohorts for Comparative Effectiveness Research

By

Harvey Hyman, PhD
University of South Florida
HymanPhD@gmail.com

Warren Fridy III, MS
Fridy Enterprises
Warren@fridyenterprises.com

Abstract

This paper describes an experiment performed on a medical record data set, using an information retrieval (IR) tool that applies the techniques of exploration and learning, to assist a researcher in identifying the most relevant cohorts. The paper presents some brief background on exploration and learning, how they are incorporated in the IR tool, and an instantiation of exploration and learning used for selecting cohorts for a research population. The research problem addressed in this paper is the TREC 2012 Medical Track task: How to provide content-based access to free-text fields of electronic medical records? The stated goal of the task is to "find a population over which comparative effectiveness studies can be done."

Report Documentation Page	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE **NOV 2012**	2. REPORT TYPE	3. DATES COVERED **00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE **Using Exploration and Learning for Medical Records Search: An Experiment in Identifying Cohorts for Comparative Effectiveness Research**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of South Florida,4202 E. Fowler Avenue,Tampa,FL,33620**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

Presented at the Twenty-First Text REtrieval Conference (TREC 2012) held in Gaithersburg, Maryland, November 6-9, 2012. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA). U.S. Government or Federal Rights License

14. ABSTRACT

This paper describes an experiment performed on a medical record data set, using an information retrieval (IR) tool that applies the techniques of exploration and learning, to assist a researcher in identifying the most relevant cohorts. The paper presents some brief background on exploration and learning, how they are incorporated in the IR tool, and an instantiation of exploration and learning used for selecting cohorts for a research population. The research problem addressed in this paper is the TREC 2012 Medical Track task: How to provide content-based access to free-text fields of electronic medical records? The stated goal of the task is to ?find a population over which comparative effectiveness studies can be done.?

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**7**	

Introduction

The problem presented here regarding how to identify cohorts from a collection of electronic medical records is an example of an information retrieval (IR) problem which relies on techniques used for extracting a maximum of relevant documents and a minimum of non-relevant documents. In typical IR problems the motivation is to reduce the time and cost of human review of the extracted collection. In this case, the motivation is to provide the best cohort population so that the ensuing research study will be useful and valid.

We employ a manual approach supported by an automated tool to address the constraint of content-based, free-text fields by creating an artifact to support the researcher in exploring a corpus of items and facilitating examining and scrutinizing. This supports user acquisition of contextual knowledge about the collection. The tool in this case has been adapted from an IR tool previously deployed for eDiscovery and presented at the TREC Conference 2011 Legal Track (Hyman and Fridy III, 2011).

Exploration

The concept of exploration has been associated with learning (Berlyne, 1963; March, 1993; familiarization (Barnett, 1963), and information search (Debowski et al, 2001). In fact work done by Berlyn in the 1960s classifies exploration as a "fundamental human activity" (Demangeot and Broderick, 2010).

Exploration is seen as a behavior motivated by curiosity. Exploration that is goal directed is classified as extrinsic (Berlyn, 1960). Extrinsic exploration typically has a specific task purpose, whereas intrinsic exploration is motivated by learning (Berlyn, 1960; Demangeot and Broderick, 2010). Kaplan and Kaplan, 1982, argue that exploration arises from our need to make sense of our environment. March, 1991, writes about exploration and exploitation. He views exploration and exploitation as competing tensions in organizational learning.

Berlyn, 1963, suggests that specific exploration is a means of satisfying curiosity. The goals of exploration as a means for making sense of our environment and satisfying curiosity are represented in the problem domain of information retrieval and in this task of cohort identification. Debowski et al, 2001, view exploratory search as a "screening process," and that exploration identifies items “to become the focus of attention.” They suggest that exploration leads to learning through the examining and scrutinizing of items.

The experiment reported in this paper presents an instance of *exploration* as a technique implemented through an IR tool as a method for identifying cohorts for a population.

Learning

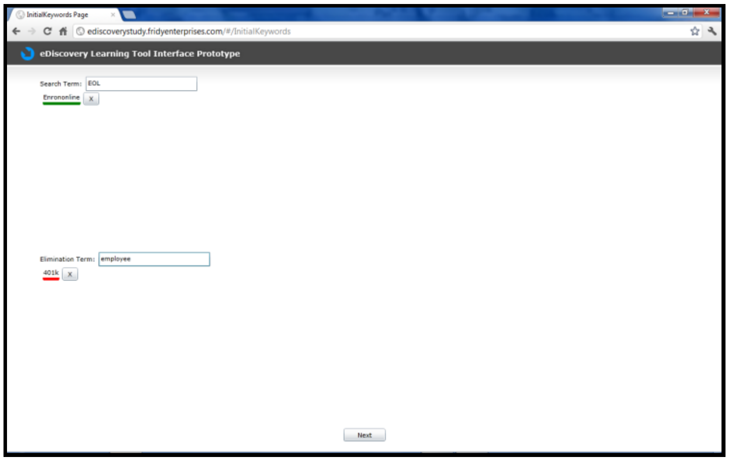
“The search for information is often a cyclical, exploratory process” (Debowski et al, 2001). Search has also been compared to problem solving techniques similar to foraging (Hills et al, 2010). Hills et al, characterize problem solving itself as a search process. The decision regarding when to exploit – stay with the current position or strategy, versus when to explore – move on to a new search or location is a trade-off that has been studied in problem solving and learning (Robbins, 1952; March 1991; Hills et al, 2010). This is especially true in the domain of content-based IR where the search can be very complex in terms of strategy and structure (Debrowski et al, 2001).

The learning process supported by the artifact allows the researcher to acquire knowledge about the records in the collection and use that knowledge to gain insight for identifying the best cohorts. In this experiment we test whether the acquired knowledge about the corpus can be effectively exploited by presenting ad hoc, iterative retrieval results to the user. An assumption herein is that the user can assess the results and adjust the search structure to improve the retrieval result – in this case identify better cohorts.

The goal of the artifact is to address the gap in electronic search identified by Dembroski et al, 2001 as; “not highly informative regarding the effectiveness of strategies.” They suggest that in order to achieve successful retrieval, the search structure and alternative strategies must be continually evaluated. We address this by using learning and iterative feedback. The context and richness discovered through exploration is applied to a corpus through an iterative learning process supported by the tool.

The Artifact

The artifact in this case is an automated tool that extracts documents from a collection that meet user criteria. Figure-1 is an example of the User Input Screen. This screen accepts the user's criteria for identifying a relevant record. The tool accepts inclusive and exclusive criteria. Figure-2 is an example of the Retrieval Screen. This screen presents the user with a sample of the extracted collection. This sample represents the content of the extraction produced by the user's criteria. The user may set a threshold for the sample. In this case we used a sample of 10 documents per extraction. The user may select on any record in the left column and view the record in the right column. After the sample has been exhausted, the user may create changes to the search criteria and the tool will present a new sample of extractions for the user to explore. The user may set a threshold for precision or a fixed number of iterations. In this case we used a fixed number of 10 iterations. The goal of the tool is to provide the user with insight into the nature of the collection and the content of the individual records through an iterative method using exploration and learning – for identifying cohorts.


InitialKeywords Page
ediscoverystudy.fridyenterprises.com/#/InitialKeywords
eDiscovery Learning Tool Interface Prototype
X
X
Next

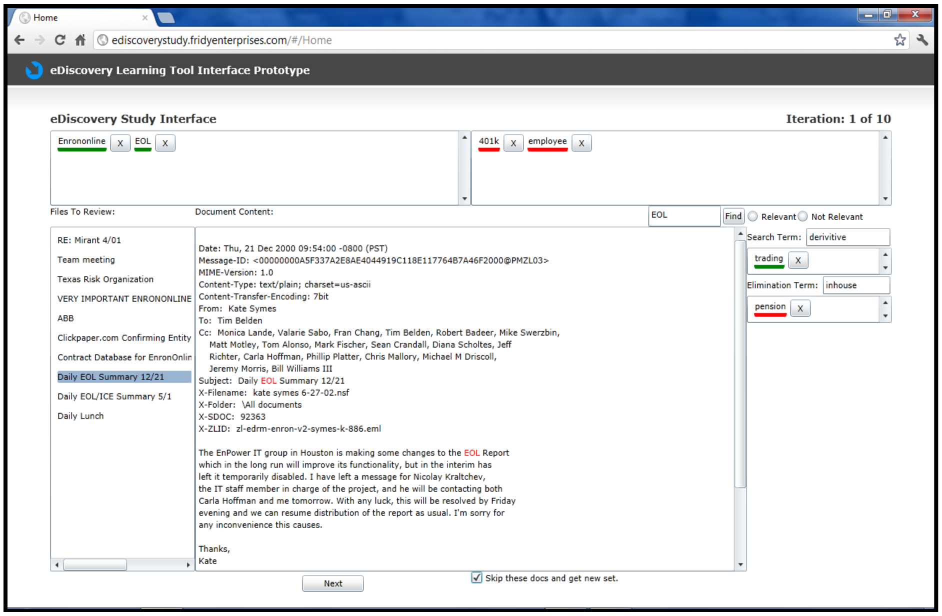
Search Term: EOL

Enrononline

Elimination Term: employee

401k

Figure-1, Source: eDiscovery Learning Tool, Fridy Enterprises


Home
ediscoverystudy.fridyenterprises.com/#/Home
eDiscovery Learning Tool Interface Prototype
Find
Next

eDiscovery Study Interface

Iteration: 1 of 10

Enrononline X EOL X

401k X employee X

Files To Review:

- RE: Mirant 4/01
- Team meeting
- Texas Risk Organization
- VERY IMPORTANT ENRONONLINE
- ABB
- Clickpaper.com Confirming Entity
- Contract Database for EnronOnlin
- Daily EOL Summary 12/21
- Daily EOL/ICE Summary 5/1
- Daily Lunch

Document Content: EOL

○ Relevant ○ Not Relevant

Search Term: derivitive

trading X

Elimination Term: inhouse

pension X

```
Date: Thu, 21 Dec 2000 09:54:00 -0800 (PST)
Message-ID: <00000000A5F337A2E8AE4044919C118E117764B7A46F2000@PMZL03>
MIME-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
From:  Kate Symes
To:  Tim Belden
Cc:  Monica Lande, Valarie Sabo, Fran Chang, Tim Belden, Robert Badeer, Mike Swerzbin,
   Matt Motley, Tom Alonso, Mark Fischer, Sean Crandall, Diana Scholtes, Jeff
   Richter, Carla Hoffman, Phillip Platter, Chris Mallory, Michael M Driscoll,
   Jeremy Morris, Bill Williams III
Subject:  Daily EOL Summary 12/21
X-Filename:  kate symes 6-27-02.nsf
X-Folder:  \All documents
X-SDOC:  92363
X-ZLID:  zl-edrm-enron-v2-symes-k-886.eml

The EnPower IT group in Houston is making some changes to the EOL Report
which in the long run will improve its functionality, but in the interim has
left it temporarily disabled. I have left a message for Nicolay Kraltchev,
the IT staff member in charge of the project, and he will be contacting both
Carla Hoffman and me tomorrow. With any luck, this will be resolved by Friday
evening and we can resume distribution of the report as usual. I'm sorry for
any inconvenience this causes.

Thanks,
Kate
```

☑ Skip these docs and get new set.

Figure-2, Source: eDiscovery Learning Tool, Fridy Enterprises

Experiment

We began with a non-function word approach. Non-function words are words such as: the, as, of, or, etc. These words serve no content purpose and provide no insight into the task. In this case we had no prior theory or knowledge about the collection. Therefore, in the absence of a specific theory to act upon or a known search strategy based on the circumstances, a non-function word approach makes the most logical sense. It provides the best possible point of view to start from because we are using the requestor's own words and terminologies. Once we entered the initial search criteria we then made adjustments based on the samples produced by the tool. We ran 10 iterations before finalizing our cohort selection.

Notable Results

We discovered that any requests for documents seeking a particular age or age group needs to be structured by a specific search for the term age and a window span of 4 characters past the term. We also discovered medical codes within the visit records. When applying a standard medical context "______" as a wild card for interpreting the codes, a cohort search to acquire documented visits based on the medical codes themselves may be structured.

Limitations

We had difficulty handling files without visit IDs. Ultimately, we were not able to resolve this issue and had to ignore those records. This reduced our effectiveness because it left us with records within the collection that we were not able to explore thoroughly.

We also had difficulty exploiting the full power of using the diagnostic codes due to our lack of experience in this domain. This has more to do with the focus of our tool. It is designed to be used to support a user who is a domain expert or a user who has a targeted idea of the direction of the search and a structure for implementing it.

Conclusion

Final results are still pending so we are unable to report the efficacy of our tool until after the conference. However, the main purpose of the discussion presented here has been to describe how exploration and learning can be instantiated in an automated tool to assist a researcher in identifying a cohort population. We welcome feedback and suggestions about our work presented.

References

Barnett, S., A., A Study in Behavior. London: Methuen (1963).

Berlyne, D., E., Conflict, Arousal and Curiosity, New York: McGraw Hill (1960).

Berlyne, D., E., "Motivational Problems Raised by Exploratory and Epistemic Behavior," Psychology: A Study of Science, Vol. 5, pp. 284-364, New York: McGraw Hill (1963).

Debowski, S., Wood, R., E., Bandura, A., "Impact of Guided Exploration and Enactive Exploration on Self-Regulatory Mechanisms and Information Acquisition Through Electronic Search," *Journal of Applied Psychology*, Vol. 86, No.6, (2001).

Demangeot, C., Broderick, A., J., "Exploration and Its Manifestations in the Context of Online Shopping," *Journal of Marketing Management*, Vol. 26, No. 13 – 14, (December, 2010).

Hills, T., T., "The Central Executive as a Search Process: Priming Exploration and Exploitation Across Domains," *Journal of Experimental Psychology*, Vol. 139, No. 4, (2010).

Hyman, H., S., Fridy, W., "Modeling Concept and Context to Improve Performance in eDiscovery," *NIST Special Publication, Proceedings: Text Retrieval Conference (TREC) 2011*.

Kaplan, S., Kaplan, R., Cognition and Environment. New York: Praeger (1982).

March, J., G., "Exploration and Exploitation in Organizational Learning," *Organizational Science*, 2(1), (1991).